WHAT IS CLAIMED IS:

1. A method for creating a structured document, wherein a structured document includes a plurality of content elements wrapped in pairs of hierarchically nested tags, comprising:

parsing a document of a particular type containing content into a plurality of content elements; and

for a selected content element, suggesting an optimal tag according to a tag suggestion procedure;

wherein the tag suggestion procedure comprises:

providing sample data in the form of structured sample documents;

analyzing patterns in the sample data to derive a set of tag suggestions;

deriving a set of candidate tags from the set of tag suggestions for the selected content element; and

evaluating the set of candidate tags according to tag suggestion criteria to determine an optimal tag for the selected content element.

2. The method of claim 1, wherein the tag suggestion criteria comprises satisfying a similarity function.

3. The method of claim 1, wherein the set of tag suggestions are generated during creation of the structured document.

4. The method of claim 1, wherein the set of tag suggestions are generated prior to creation of the structured document.

5. The method of claim 1, wherein the structured sample document comprises an XML document having a DTD associated with it.

6. The method of claim 1, wherein the set of tag suggestions includes tree patterns of tags.

7. The method of claim 1, wherein the optimal tag maximizes a similarity function with patterns found in the sample data.

8. The method of claim 6, wherein the tag suggestion criteria comprises balancing size of tree patterns of tags and frequency of occurrence of tree patterns of tags in the sample data.

9. The method of claim 1, wherein the set of tag suggestions includes a set of tree patterns of tags $t_i \in T$, and a set C of candidates is a set of all patterns in T with all their prefixes, $C = \{c \mid c \text{ is a prefix of } t_i \in T\}$;

wherein a similarity function between a candidate $c \in C$ and a tree pattern $t_i \in T$ satisfies: $sim(c, t_i) = |c| / |t_i|$, if c is a tree-prefix of t_i;

$sim(c, t_i) = 0$, otherwise; and

wherein the optimal tag comprises a context-free candidate $c \in C$ that maximizes an aggregate similarity measure $SIM(c,T)$, where $SIM(c,T) = \sum_{t_i \in T} sim(c,t_i) \cdot pr_i$.

10. The method of claim 9, wherein a candidate set in context t_{cxt} is defined as $C(t_{ctx}) = \{ c \in C \mid t_{cxt} \text{ is a prefix of } c \}$; and

wherein the optimal tag comprises a context-aware candidate $c \in C$ that maximizes an aggregate similarity measure $SIM(c,T)$, where $SIM(c,T) = \sum_{t_i \in T} sim(c,t_i) \cdot pr_i$.

11. A method for authoring of a structured document, wherein a structured document comprises a plurality of content elements wrapped in pairs of tags, comprising:

generating content elements wrapped in pairs of tags; and

for a selected tag, suggesting an optimal content fragment according to a contentt suggestion procedure;

wherein the content suggestion procedure comprises:

providing a sample structured document;

deriving a set of content fragments from the sample structured document;

evaluating the set of content fragments according to a content fragment suggestion criteria to determine an optimal content fragment suggestion for the tag, wherein the optimal content fragment suggestion is the most probable content fragment for the selected tag.

12. The method of claim 11, further comprising assigning a score to each content fragment in the set of content fragments, wherein the score is a ratio of number of occurrences of the content fragment under the selected tag and number of occurrences of the selected tag in the sample structured document.

13. The method of claim 12, wherein the optimal content fragment suggestion is the content fragment with the highest score.

14. The method of claim 12, further comprising assigning a context to each content fragment in the set of content fragments, wherein context comprises the structural context of the tag surrounding the content fragment.

15. The method of claim 12, wherein the optimal content fragment suggestion is the content fragment with the highest score greater than a threshold value.

16. The method of claim 14, wherein each content fragment is referenced by a partial path from the sample structured document root and the context comprises the partial path of the content fragment in the sample structured document.

17. The method of claim 11, further comprising:

selecting a small linguistic unit within each content fragment in the set of content fragments; and

assigning a score to the small linguistic unit, wherein the score is a ratio of number of occurrences of the linguistic unit under the selected tag and number of occurrences of the selected tag in the sample structured document.

18. The method of claim 17, wherein the small linguistic unit is a word, a phrase or a sentence.

19 The method of claim 14, wherein the context of each content fragment in the set of content fragments comprises the structural tree around the tag surrounding the content fragment.

20. The method of claim 1, wherein content comprises text.